CE888 Assignment 2 (21a3)

February 16, 2022

Set by:	Dr Ana Matran-Fernandez (`amatra@essex.ac.uk`)
Submission deadline:	Week 30 (see FASER for exact date and time)
Feedback:	Week 33 (3 weeks from submission deadline)
Submission mode:	Electronic via FASER (see Section "Submission" below)

1 Assignment objectives

This document specifies the coursework assignment to be submitted by students taking CE888.

The main aims of this assignment are:

1. To present models and data interpretations.

2. To create a complete data science project and make it available in an open-access repository.

3. To summarise your findings in a report.

4. To provide recommendations based on your findings.

You should incorporate the feedback from Assignment 1 into your work here. In order to avoid self-plagiarism, you will need to paraphrase the content from your Assignment 1 if needed (see Section 4).

2 The Assignment

The manager of the company that asked you for advice on their problem has given the green light for a more detailed execution of your plan. After providing you with feedback on your original report, she has asked you to incorporate it into your work, carry out the modelling part of the project, and report your main findings and recommendations for the company.

2.1 The Report

The report should be written using an adequate level of English and include the following sections (use the template provided on Moodle):

1. **Title**: Make sure the title of your report is descriptive of your work. You can be creative with this. Imagine this is going to be read by the company manager. Do not use "Project 1/2/3/Reassessment/Assignment 1/2" as your title.

2. **Executive summary**: Provide a short description of your work, summarising your main findings. A good summary should include a description of the problem as you understand it and its significance, a summary of the methods and results, and a short conclusion.

   The executive summary should not be longer than 250 words. It should not include references.

3. **Introduction**: Explain the purpose of your work and motivates it – why is what you are doing important? This section should include references to show why what you are doing is relevant and to back up any claims you make. We expect at least 5 good-quality references (i.e., peer-reviewed articles and/or books; not Wikipedia, YouTube or general websites).

   Use a maximum of 600 words for this section (excluding references). Normally there should be no figures or tables in the Introduction.

4. **Data**: Should be divided into subsections, one for each dataset used.

   Describe the dataset/s you have used, including how the data was collected (or generated). For each dataset, you must give information about its size, type of features, all the preprocessing steps that were done, any train/validation/test splits, etc. If you chose the dataset/s, justify why they are appropriate for the project you are doing.

   This is not an exhaustive list, and you should give more information as appropriate.

   The word limit for this section is 300 words/dataset, excluding references, figures, and tables. Note that this is shorter than in the previous assignment.

5. **Methodology**: Describe the methodology you followed in detail, so that a data analyst can replicate your steps without looking at your code. This includes hyper-parameters (if different from default values in the library you used). Do not include code in either text, figure, or table forms. The code is evaluated separately. This is a description of the methods in scientific English.

   Think of this section as an updated section of the Methodology that you wrote for Assignment 1, taking into account any feedback you received from us.

   The word limit for this section is 600 words, excluding references, figures, and tables and their captions. Use subsections as needed.

6. **Results**: Use figures and tables to show your results, including intermediate results you may have obtained throughout the process. Check examples of papers from others (e.g., here) to describe the results.

7. **Discussion**: What are insights that you can extract from your results? Compare your results from others if applicable. What is better/worse in your approach? Are there any limitations to the methodology you followed?

   The word limit for the Results and Discussion sections together is 1000 words, excluding references, figures, tables, and captions.

8. **Conclusions/Recommendations**: A couple of short paragraphs describing any concluding remarks you might have, including potential avenues for future work/improvements and recommendations for the company.

   The word limit for this section is 500 words.

2.2 The Code

All the code should be in GitHub (ideally, using .py files for the main code, and some illustrative examples and data exploration on Jupyter Notebooks). The notebooks should not be used to store all the code, just for examples, properly importing functions from Python files.

The GitHub repository should include:

- a README file with links/instructions to download the datasets and a description of the repository and how to use/run the code.

- The code that you used to carry out the project. The code should match your Methodology section and be well documented. If you use Jupyter Notebooks, there should be comments on the findings and not a stream of figures/plots with no justification of why they're informative.

3 Do's and don't's

- This is an individual project and you must work on it by yourself.

- DO read this document twice and check the assignment template and the marking scheme.

- DO read your previous report and the feedback received.

- DO start a thread on the Moodle forum if you have problems/questions about the assignment.

- DO save figures properly from Python and include them in the report. I recommend saving them in pdf format.

- DO ensure that each table and figure has an appropriate caption describing it. Refer to them by their number, and not by "figure/table above/below". Tables and figures should be placed at the top or bottom of the page they are in. All tables and figures should be referred to in the text.

- DO use functions and comments in your code.

- DO write comments and observations if you are using Jupyter Notebooks. If you are using Python scripts, make sure they are well commented too.

- DO paraphrase your previous assignment (if needed) to avoid self-plagiarism.

- DO NOT wait until the last week to get started on the assignment.

- DO NOT copy and paste from other sources (with or without referencing) — this is plagiarism.

- DO NOT copy text from other sources and replace random words — this is also plagiarism (with or without referencing). You must paraphrase the text (and add a reference to it).

- DO NOT include screenshots of your code or code outputs in the report. Any numerical data that you include should be in a suitable graphical or tabular form. You should not include any numerical data that is not relevant to your discussion (do not trivially copy/paste raw output produced by your code).

- DO NOT write your name on the report: use your registration number.

4   A note on paraphrasing

Paraphrasing is more than changing some words in the text (this makes it unreadable and will penalise you). For example, referring to a "random forest" as a "random collection of trees" is not scientific and it does not make sense (and it has been done by students previously!). Other real-life examples include replacing "cross-validation" with "cross-approval", and "deep neural network" with "profound neural organization".

Use your head: read the text, think about the idea you want to convey, and write it down in your own words without looking at the original source. Make sure you add references to the original source/s.

5   Submission

Your work must be submitted to FASER by the deadline given on the system. No other mode of submission is acceptable. Do not wait until close to the deadline to make your submission. Difficulties with the submission system will not be accepted as an excuse for a missing/late submission.

You must submit a report in PDF format (no doc, docx, etc.), adhering to the CE888 Assignment 2 template available on Moodle.

To get the word count per section, you can use: `https://app.uio.no/ifi/texcount/online.php`

6   Marking criteria

The assignment is worth 70% of the module mark and will be assessed according to the marking scheme available on Moodle.